Em um cenário em constante evolução, o tema da detecção de fraudes em pagamentos bancários continua a despertar preocupações e ações. A cada ano, a fraude se torna uma indústria que movimenta bilhões de dólares, impactando setores diversos, como comércio eletrônico, saúde e sistemas de pagamento.

A detecção de fraudes se refere à identificação e prevenção de atividades fraudulentas, que envolvem ações enganosas ou ilegais com o objetivo de obter vantagens financeiras, informações pessoais ou outros ganhos indevidos.

Uma abordagem antiga para combater fraudes envolvia o uso de sistemas de regras, que eram definidas manualmente e usadas para tomar decisões sobre a autenticidade de transações, atividades ou comportamentos suspeitos.  
No entanto, essa abordagem apresenta diversas limitações, entre elas, a falta de adaptabilidade, que permite que novos tipos de fraude passem despercebidos até que as regras sejam atualizadas manualmente. Além disso, há o desafio do aumento do volume e complexidade dos dados, tornando a gestão por meio de regras estáticas uma tarefa difícil, e também vale ressaltar que, a configuração inadequada das regras pode resultar em ocorrências de "falsos positivos" ou "falsos negativos". Essa situação pode causar tanto inconvenientes para os clientes legítimos quanto perdas financeiras para a organização.

Nesse contexto, o potencial do aprendizado de máquina emerge como uma ferramenta poderosa para aprimorar a precisão e a eficácia da detecção de fraudes. Esses modelos podem analisar grandes volumes de dados, identificar padrões sutis e aprender com novos exemplos de fraudes à medida que surgem, tornando-se mais adaptáveis e eficazes em comparação com sistemas de regras estáticas, além de reduzir os falsos positivos e negativos, com isso, os bancos podem reduzir riscos financeiros, aumentar a confiabilidade do sistema de pagamento e manter a confiança dos clientes.

Em suma, enfrentar o desafio das fraudes em pagamentos bancários exige uma abordagem multidimensional que incorpore tanto o conhecimento especializado presente nas regras quanto a capacidade de aprendizado e adaptação dos modelos de aprendizado de máquina. Com a tecnologia evoluindo constantemente, a detecção de fraudes está se tornando mais precisa e ágil, proporcionando um ambiente financeiro mais seguro e confiável para todas as partes envolvidas

O uso de algoritmos de detecção de anomalias, classificação e agrupamento de dados, redes neurais e outras técnicas avançadas de aprendizado de máquina possibilita uma detecção mais precisa e em tempo real de atividades fraudulentas, além de reduzir os falsos positivos e negativos.

Esses métodos podem analisar grandes volumes de dados, identificar padrões sutis e aprender com novos exemplos de fraudes à medida que surgem, tornando-se mais adaptáveis e eficazes em comparação com sistemas de regras estáticas

Por outro lado, o aprendizado de máquina é probabilístico em vez de determinístico. Ele utiliza modelos estatísticos para examinar resultados passados e anomalias a fim de prever resultados futuros. Um sistema de aprendizado de máquina pode aprender, prever e tomar decisões sem ser programado explicitamente, um sistema de aprendizado de máquina pode distinguir as características de compras fraudulentas das legítimas. O aprendizado de máquina frequentemente é implantado como parte de sistemas automatizados de triagem de fraudes, identificando transações de alto risco, contas e logins suspeitos para evitar fraudes de pagamento, abuso de contas, abuso de conteúdo e tomada de contas. O aprendizado de máquina pode substituir até mesmo conjuntos de regras mais complexos e produzir maior precisão, menos resultados falsos positivos e economia por meio da automação

Para superar essas limitações, muitas organizações estão adotando abordagens mais avançadas para a detecção de fraudes. Uma dessas abordagens é a utilização de técnicas de aprendizado de máquina e inteligência artificial. Esses métodos podem analisar grandes volumes de dados, identificar padrões sutis e aprender com novos exemplos de fraudes à medida que surgem, tornando-se mais adaptáveis e eficazes em comparação com sistemas de regras estáticas.

O uso de algoritmos de detecção de anomalias, classificação e agrupamento de dados, redes neurais e outras técnicas avançadas de aprendizado de máquina possibilita uma detecção mais precisa e em tempo real de atividades fraudulentas, além de reduzir os falsos positivos e negativos.

Ao aproveitar a capacidade de adaptabilidade e aprendizado contínuo dos modelos de IA, os bancos podem reduzir riscos financeiros, aumentar a confiabilidade do sistema de pagamento e manter a confiança dos clientes.

Em suma, enfrentar o desafio das fraudes em pagamentos bancários exige uma abordagem multidimensional que incorpore tanto o conhecimento especializado presente nas regras quanto a capacidade de aprendizado e adaptação dos modelos de aprendizado de máquina. Com a tecnologia evoluindo constantemente, a detecção de fraudes está se tornando mais precisa e ágil, proporcionando um ambiente financeiro mais seguro e confiável para todas as partes envolvidas

A

Nesse contexto, o potencial do aprendizado de máquina emerge como uma ferramenta poderosa para aprimorar a precisão e a eficácia da detecção de fraudes. A combinação de abordagens, como a fusão de regras e algoritmos de aprendizado de máquina, representa uma estratégia promissora para atingir resultados mais confiáveis e seguros. Ao aproveitar a capacidade de adaptabilidade e aprendizado contínuo dos modelos de IA, os bancos podem reduzir riscos financeiros, aumentar a confiabilidade do sistema de pagamento e manter a confiança dos clientes.

Em suma, enfrentar o desafio das fraudes em pagamentos bancários exige uma abordagem multidimensional que incorpore tanto o conhecimento especializado presente nas regras quanto a capacidade de aprendizado e adaptação dos modelos de aprendizado de máquina. Com a tecnologia evoluindo constantemente, a detecção de fraudes está se tornando mais precisa e ágil, proporcionando um ambiente financeiro mais seguro e confiável para todas as partes envolvidas.

O uso de aprendizado de máquina em detecção de fraude em comparação com regras tradicionais.

Em diversas indústrias, o aprendizado de máquina está substituindo soluções antigas que não conseguem acompanhar o ritmo ou entregar a mesma qualidade de resultados.

Na detecção de fraudes, a abordagem ultrapassada para combater fraudes é o uso de sistemas de regras atualizadas manualmente, que dependem de declarações condicionais para tomar decisões. O sistema percorre as regras, uma por uma, e se determinar que alguma regra foi ativada, ele tomará a ação apropriada e ignorará todas as outras regras.

Por outro lado, o aprendizado de máquina é probabilístico em vez de determinístico. Ele utiliza modelos estatísticos para examinar resultados passados e anomalias a fim de prever resultados futuros. Um sistema de aprendizado de máquina pode aprender, prever e tomar decisões sem ser programado explicitamente, um sistema de aprendizado de máquina pode distinguir as características de compras fraudulentas das legítimas. O aprendizado de máquina frequentemente é implantado como parte de sistemas automatizados de triagem de fraudes, identificando transações de alto risco, contas e logins suspeitos para evitar fraudes de pagamento, abuso de contas, abuso de conteúdo e tomada de contas. O aprendizado de máquina pode substituir até mesmo conjuntos de regras mais complexos e produzir maior precisão, menos resultados falsos positivos e economia por meio da automação.

A detecção de fraudes é uma área crítica em diversos setores, como finanças, comércio eletrônico e segurança cibernética. Ela se refere à identificação e prevenção de atividades fraudulentas, que envolvem ações enganosas ou ilegais com o objetivo de obter vantagens financeiras, informações pessoais ou outros ganhos indevidos. Uma abordagem antiga para combater fraudes envolvia o uso de sistemas de regras, que eram definidas manualmente e usadas para tomar decisões sobre a autenticidade de transações, atividades ou comportamentos suspeitos.

Esses sistemas de regras funcionam por meio de declarações condicionais, essas regras eram criadas com base na experiência passada e no conhecimento disponível sobre os padrões de fraude.

No entanto, essa abordagem apresenta várias limitações, como falta de adpatabilidade, que permite que novos tipos de fraude passem despercebidos até que as regras sejam atualizadas manualmente, alto volume e complexidade dos dados, já que é dificil gerencia-los por meio de regras estáticas, e também, dependendo da configuração das regras, pode haver casos de "falsos positivos" (transações legítimas marcadas como suspeitas) ou "falsos negativos" (transações fraudulentas que não são detectadas). Isso pode levar a inconveniências para os clientes legítimos ou a perdas financeiras para a organização.

1. **Rigidez e Falta de Adaptabilidade**: As fraudes evoluem constantemente e novos padrões de comportamento fraudulento emergem com o tempo. As regras estáticas não conseguem se adaptar rapidamente a essas mudanças, o que permite que novos tipos de fraude passem despercebidos até que as regras sejam atualizadas manualmente.
2. **Volume e Complexidade dos Dados**: Com o aumento do volume de dados, as regras podem se tornar excessivamente complexas e difíceis de gerenciar. Também é desafiador capturar todas as nuances e variações dos comportamentos fraudulentos por meio de regras estáticas.
3. **Taxa de Falsos Positivos e Negativos**: Dependendo da configuração das regras, pode haver casos de "falsos positivos" (transações legítimas marcadas como suspeitas) ou "falsos negativos" (transações fraudulentas que não são detectadas). Isso pode levar a inconveniências para os clientes legítimos ou a perdas financeiras para a organização.
4. **Mudança de Comportamento dos Usuários**: Às vezes, o comportamento legítimo dos usuários pode ser interpretado erroneamente como fraude devido à aplicação rígida das regras.

Para superar essas limitações, muitas organizações estão adotando abordagens mais avançadas para a detecção de fraudes. Uma dessas abordagens é a utilização de técnicas de aprendizado de máquina e inteligência artificial. Esses métodos podem analisar grandes volumes de dados, identificar padrões sutis e aprender com novos exemplos de fraudes à medida que surgem, tornando-se mais adaptáveis e eficazes em comparação com sistemas de regras estáticas.

O uso de algoritmos de detecção de anomalias, classificação e agrupamento de dados, redes neurais e outras técnicas avançadas de aprendizado de máquina possibilita uma detecção mais precisa e em tempo real de atividades fraudulentas, além de reduzir os falsos positivos e negativos.

Em resumo, embora a abordagem de sistemas de regras manuais tenha sido uma estratégia comum no passado para combater fraudes, as limitações inerentes a essa abordagem impulsionaram a adoção de métodos mais avançados, como o uso de aprendizado de máquina e inteligência artificial, para melhorar a detecção de fraudes de forma mais eficaz e adaptável.

Introdução

Em um cenário em constante evolução, o tema da detecção de fraudes em pagamentos bancários continua a despertar preocupações e ações. A cada ano, a fraude se torna uma indústria que movimenta bilhões de dólares, impactando setores diversos, como comércio eletrônico, saúde e sistemas de pagamento.

A detecção de fraudes se refere à identificação e prevenção de atividades fraudulentas, que envolvem ações enganosas ou ilegais com o objetivo de obter vantagens financeiras, informações pessoais ou outros ganhos indevidos.

Uma abordagem antiga para combater fraudes envolvia o uso de sistemas de regras, que eram definidas manualmente e usadas para tomar decisões sobre a autenticidade de transações, atividades ou comportamentos suspeitos.  
No entanto, essa abordagem apresenta diversas limitações, entre elas, a falta de adaptabilidade, que permite que novos tipos de fraude passem despercebidos até que as regras sejam atualizadas manualmente. Além disso, há o desafio do aumento do volume e complexidade dos dados, tornando a gestão por meio de regras estáticas uma tarefa difícil, e também vale ressaltar que, a configuração inadequada das regras pode resultar em ocorrências de "falsos positivos" ou "falsos negativos". Essa situação pode causar tanto inconvenientes para os clientes legítimos quanto perdas financeiras para a organização.

Nesse contexto, o potencial do aprendizado de máquina emerge como uma ferramenta poderosa para aprimorar a precisão e a eficácia da detecção de fraudes. Esses modelos podem analisar grandes volumes de dados, identificar padrões sutis e aprender com novos exemplos de fraudes à medida que surgem, tornando-se mais adaptáveis e eficazes em comparação com sistemas de regras estáticas, além de reduzir os falsos positivos e negativos, com isso, os bancos podem reduzir riscos financeiros, aumentar a confiabilidade do sistema de pagamento e manter a confiança dos clientes.

<https://www.insper.edu.br/noticias/crescimento-de-golpes-no-comercio-eletronico-impulsiona-mercado-de-prevencao/>

<https://resources.sift.com/ebook/machine-learning-fraud-prevention-whats-next/>

**Descrição do Problema:**

O problema abordado pela pesquisa é a detecção de fraudes em pagamentos bancários. A fraude nesse contexto envolve atividades fraudulentas que visam obter vantagens financeiras ou informações pessoais de maneira enganosa ou ilegal. Essas atividades podem causar prejuízos significativos para diferentes setores, como comércio eletrônico, saúde e sistemas de pagamento.

**Justificativa e Motivação:** A detecção de fraudes em pagamentos é uma questão crítica devido ao aumento constante das atividades fraudulentas, que se transformaram em uma indústria bilionária. As abordagens tradicionais, como sistemas de regras manuais, têm limitações consideráveis, incluindo falta de adaptabilidade a novos tipos de fraude, dificuldade na gestão de dados complexos e configuração inadequada de regras, levando a "falsos positivos" e "falsos negativos". Tais problemas resultam em inconvenientes para clientes legítimos e perdas financeiras para as organizações.

**Objetivos:** O principal objetivo da pesquisa é melhorar a detecção de fraudes em pagamentos bancários usando técnicas de Inteligência Artificial (IA) e Aprendizado de Máquina (AM). Isso será alcançado através dos seguintes subobjetivos:

1. Desenvolver modelos de aprendizado de máquina capazes de analisar grandes volumes de dados de transações bancárias.
2. Identificar padrões sutis e complexos que caracterizam atividades fraudulentas.
3. Criar sistemas adaptáveis que possam aprender com novos exemplos de fraudes à medida que surgem.
4. Reduzir os casos de "falsos positivos" e "falsos negativos" para melhorar a eficácia do sistema de detecção.
5. Contribuir para a confiabilidade do sistema de pagamento e manter a confiança dos clientes.

**Solução Proposta:** A solução proposta envolve a aplicação de técnicas de Inteligência Artificial e Aprendizado de Máquina para aprimorar a detecção de fraudes em pagamentos bancários. Serão utilizados algoritmos de Aprendizado de Máquina, como Random Forest, Support Vector Machines (SVM), Redes Neurais e algoritmos de Clusterização para cumprir os objetivos estabelecidos.

A tarefa de Aprendizado de Máquina consistirá em treinar esses modelos com um conjunto diversificado de dados de transações bancárias, incluindo informações sobre os clientes, histórico de transações e padrões comportamentais. Os modelos serão ajustados para reconhecer padrões específicos associados a fraudes, permitindo a identificação de atividades suspeitas.

Além disso, será implementado um sistema de aprendizado contínuo, onde os modelos serão atualizados com novos exemplos de fraudes à medida que são descobertos. Isso tornará os modelos mais adaptáveis a novos métodos de fraude.

**Conclusão:** A aplicação de Inteligência Artificial e Aprendizado de Máquina na detecção de fraudes em pagamentos bancários oferece uma solução inovadora e eficaz para combater as atividades fraudulentas em constante evolução. Essa abordagem supera as limitações dos sistemas de regras estáticas, aumenta a precisão da detecção e reduz os falsos positivos e negativos. Com isso, os bancos podem proteger seus clientes, reduzir riscos financeiros e manter a confiança em seus sistemas de pagamento.

A detecção de fraude em transações financeiras envolve a identificação de atividades suspeitas que se desviam dos padrões legítimos. Essas atividades podem causar prejuízos significativos para o sistema de pagamentos financeiros.

O desafio deste projeto será explorar um dataset de grande volume de dados e aplicar modelos de aprendizado de maquiná que possa identificar padrões e anamolias, e prever a probabilidade de fraudes em novas transacoes, classificando-as em fraudulentas ou não fraudulentas, reduzindo casos de falsos positivos e negativos, melhorando a eficacia na deteccao de transacoes bancarias fraudulentas , contribuindo para a confiabilidade do sistema bancário e manter a confiança da organização.

A solução proposta envolve a aplicação de técnicas de Inteligência Artificial e Aprendizado de Máquina para aprimorar a detecção de fraudes em pagamentos bancários. Serão utilizados algoritmos de Aprendizado de Máquina, como Random Forest, XgBoost para cumprir os objetivos estabelecidos. Estes algoritmos de aprendizado de máquina são amplamente utilizados para tarefas de detecção de fraude bancária devido à sua capacidade de lidar com dados complexos e não lineares. Eles são considerados métodos ensemble, o que significa que combinam as previsões de vários modelos individuais para melhorar o desempenho geral. O modelo Random Forest é baseado em arvores de decisao, ele cria múltiplas árvores de decisão durante o treinamento e combina suas previsões para chegar a uma decisão final. Cada árvore é treinada em uma amostra aleatória dos dados, permitindo a diversificação e reduzindo o risco de overfitting. O algoritmo também utiliza o conceito de "bagging" (bootstrap aggregating), que ajuda a reduzir a variância das previsões. E o modelo XGBoost, também é baseado em arvores de decisão, porém usa abordagem de boosting, que consiste em melhorar o desempenho de modelos de aprendizado fracos, combinando-os em um modelo forte, e pode lidar bem com desequilíbrios entre classes. Ambos os algoritmos são altamente configuráveis, permitindo ajustes para otimizar o desempenho.

A tarefa de Aprendizado de Máquina consistirá em treinar esses modelos com um conjunto diversificado de dados de transações bancárias, incluindo informações sobre os clientes, histórico de transações e

Ao refinar continuamente o modelo e mantê-lo atualizado com novos dados, as empresas podem ficar um passo à frente dos fraudadores e proteger a si mesmas e a seus clientes contra perdas financeiras.

A detecção de fraude em transações financeiras envolve a identificação de atividades suspeitas que se desviam dos padrões legítimos. Essas atividades podem causar prejuízos significativos para o sistema de pagamentos financeiros.

O desafio deste projeto será explorar um dataset de grande volume de dados e aplicar modelos de aprendizado de máquina que possam identificar padrões e anomalias, e prever a probabilidade de fraudes em novas transações, classificando-as em fraudulentas ou não fraudulentas, reduzindo casos de falsos positivos e negativos, melhorando a eficácia na detecção de transações bancárias fraudulentas, contribuindo para a confiabilidade do sistema bancário e mantendo a confiança da organização.

A solução proposta envolve a aplicação de técnicas de Inteligência Artificial e Aprendizado de Máquina para aprimorar a detecção de fraudes em pagamentos bancários. Serão utilizados algoritmos de Aprendizado de Máquina, como Random Forest e XgBoost, para cumprir os objetivos estabelecidos. Esses algoritmos de aprendizado de máquina são amplamente utilizados para tarefas de detecção de fraude bancária devido à sua capacidade de lidar com dados complexos e não lineares. Eles são considerados métodos ensemble, o que significa que combinam as previsões de vários modelos individuais para melhorar o desempenho geral.

O modelo Random Forest é baseado em árvores de decisão. Ele cria múltiplas árvores de decisão durante o treinamento e combina suas previsões para chegar a uma decisão final. Cada árvore é treinada em uma amostra aleatória dos dados, permitindo a diversificação e reduzindo o risco de overfitting. O algoritmo também utiliza o conceito de "bagging" (bootstrap aggregating), que ajuda a reduzir a variância das previsões. Já o modelo XGBoost também é baseado em árvores de decisão, porém usa uma abordagem de boosting, que consiste em melhorar o desempenho de modelos de aprendizado fracos, combinando-os em um modelo forte, e pode lidar bem com desequilíbrios entre classes. Ambos os algoritmos são altamente configuráveis, permitindo ajustes para otimizar o desempenho.

A tarefa de Aprendizado de Máquina consistirá em treinar esses modelos com um conjunto diversificado de dados de transações bancárias, incluindo informações sobre os clientes e histórico de transações. Ao refinar continuamente o modelo e mantê-lo atualizado com novos dados, as organizações podem ficar um passo à frente dos fraudadores e proteger a si mesmas e seus clientes contra perdas financeiras.

O processo envolverá algumas etapas principais como coleta dos dados, com registros históricos de transações, incluindo recursos como valor da transação e informações do cliente. Pré-processamento de dados, limpeza e transformação dos dados para garantir qualidade e compatibilidade com algoritmos de aprendizado de máquina,

Análise e Exploração dos daods

* Feature engineering
* Preparação de dados para o(s) modelo(s)
* Construção de modelo
* Validação e avaliação do modelo
* Avaliação e discussão dos resultados

Construindo um modelo de aprendizado de máquina: treinamento para detecção de fraude

Para construir um modelo de detecção de fraude, podemos empregar uma variedade de algoritmos de aprendizado de máquina, incluindo:

Regressão logística: um algoritmo de classificação binária que modela a probabilidade de uma transação ser fraudulenta com base em recursos de entrada.

Random Forest: um poderoso algoritmo de conjunto que combina várias árvores de decisão para fazer previsões. Ele pode lidar com interações de recursos complexos e detectar anomalias de forma eficaz.

Gradient Boosting: Um algoritmo de aumento que cria um modelo preditivo forte combinando iterativamente modelos fracos. É particularmente útil para lidar com conjuntos de dados desequilibrados.

Vamos dar uma olhada em um exemplo de código Python usando a popular biblioteca scikit-learn para treinar um classificador Random Forest para detecção de fraude:

Coleta dos dados

BankSim é um simulador baseado em agentes de pagamentos bancários, usando dados agregados de transações fornecidos por um banco na Espanha. Seu principal objetivo é gerar dados sintéticos para pesquisa em detecção de fraudes. Foram usadas análises estatísticas e de Rede Social (SNA) para desenvolver e calibrar o modelo, analisando as relações entre comerciantes e clientes.

Executamos o BankSim por 180 etapas (aproximadamente seis meses), várias vezes, e calibramos os parâmetros para obter uma distribuição próxima o suficiente para testes confiáveis. Coletamos vários arquivos de log e selecionamos o mais preciso. Inserimos ladrões que visam roubar em média três cartões por etapa e realizar cerca de duas transações fraudulentas por dia. Produzimos um total de 594.643 registros, sendo 587.443 pagamentos normais e 7.200 transações fraudulentas. Como esta é uma simulação randomizada, os valores, é claro, não são idênticos aos dados originais.

A modelagem e simulação baseada em agentes, ou Agent-based Modeling and Simulation (ABMS), é uma abordagem para o desenvolvimento de sistemas compostos de agentes autônomos que interagem para simular sistemas complexos, nos quais o comportamento emergente é usualmente desconhecido.

Um simulador baseado em agentes é um modelo de simulação onde entidades individuais, chamadas agentes, agem e interagem autonomamente de acordo com regras pré-definidas, permitindo a observação dos padrões emergentes resultantes de suas interações.

Os dados que serão utilizados neste projeto foram obtidos em Jul/23, na plataforma Kaggle, ~~através do link (~~[~~https://www.kaggle.com/datasets/ealaxi/banksim1~~](https://www.kaggle.com/datasets/ealaxi/banksim1)~~),~~ onde também possui um link para acesso ao artigo original sobre a geração dos dados do BankSim.   
BankSim é um simulador de pagamentos bancários, baseada em agentes, ou Agent-based Modeling, abordagem utilizada para sistemas complexos, nos quais os agentes (entidades individuais) interagem de acordo com regras pré-definidas, permitindo a observação dos padrões emergentes resultantes de suas interações. Neste caso, o simulador usa dados agregados de transações fornecidos por um banco na Espanha. Seu principal objetivo é gerar dados sintéticos para pesquisa em detecção de fraudes, combinando pagamentos normais e assinaturas de fraudes conhecidas. Foram usadas análises estatísticas e Social Network Analysis (SNA), para desenvolver e calibrar o modelo, analisando as relações entre comerciantes e clientes.

Este conjunto de dados gerado sinteticamente consiste em pagamentos de vários clientes feitos em diferentes períodos de tempo e com diferentes valores. O dataset possui um total de 594.643 registros. Onde 587.443 são transações normais e 7200 transações fraudulentas.

Os dados que serão utilizados neste projeto foram obtidos em julho de 2023, na plataforma Kaggle, onde também está disponível um link para acessar o artigo original sobre a geração dos dados do BankSim.

BankSim é um simulador de pagamentos bancários baseado em agentes, ou Agent-based Modeling em inglês, abordagem utilizada para sistemas complexos, nos quais os agentes (entidades individuais) interagem de acordo com regras pré-definidas. Isso permite observar os padrões emergentes resultantes de suas interações. Neste caso, o simulador utiliza dados agregados de transações fornecidos por um banco na Espanha. Seu principal objetivo é gerar dados sintéticos para pesquisa em detecção de fraudes, combinando pagamentos normais e assinaturas de fraudes conhecidas. Foram realizadas análises estatísticas e Análise de Rede Sociais (Social Network Analysis) para desenvolver e calibrar o modelo, analisando as relações entre comerciantes e clientes.

Este conjunto de dados gerado sinteticamente consiste em pagamentos de vários clientes feitos em diferentes períodos de tempo, e com diferentes valores. O dataset possui um total de 594.643 registros, dos quais 587.443 são transações normais e 7.200 são transações fraudulentas.

Esses dados contêm vários milhares

registros de dados transacionais cobrindo seis meses, de

novembro de 2012 até abril de 2013 restrito por código postal

localização para Madrid e Barcelona. Ou seja, esses dados são

recente o suficiente para refletir as condições atuais de pagamentos,

mas agregados para não representar um risco de um cliente específico

ponto de vista da privacidade.

Conjuntos de dados gerados pelo simulador de pagamentos BankSim

<https://www.kaggle.com/datasets/ealaxi/banksim1>

Processamento/Tratamento de Dados

Nesta seção você deve registrar os métodos e ferramentas utilizados para o tratamento dos dados, bem como os algoritmos de Inteligência Artificial e Aprendizado de Máquina desenvolvidos/utilizados para esse fim. Cada procedimento realizado deve ser descrito, de forma que outra pessoa consiga reproduzir o que você fez (princípio da *replicabilidade*). Sempre justifique as decisões tomadas no desenvolvimento do seu trabalho. Sugere-se que partes dos códigos-fonte desenvolvidos sejam apresentadas como forma de mostrar o passo a passo desta etapa.

. A análise estatística dos dados permitiria, dessa forma, a observação do impacto das variáveis selecionadas no referido risco de crédito. Tendo por base o referencial teórico e os estudos feitos por outros autores, esperava-se então, que características individuais, tais como, idade, sexo, estado civil, tempo de relacionamento com o banco, volume de investimento, existência ou histórico de restrições cadastrais, renda do proponente, natureza ocupacional fossem significativas para determinar o perfil do cliente e sua propensão à inadimplência. A Tabela 1 apresenta a estatística descritiva das variáveis tidas como significativas na explicação do risco de crédito dos clientes em análise. Destacam-se na tabela, a média, o desvio-padrão e coeficientes de assimetria e curtose de cada variável.

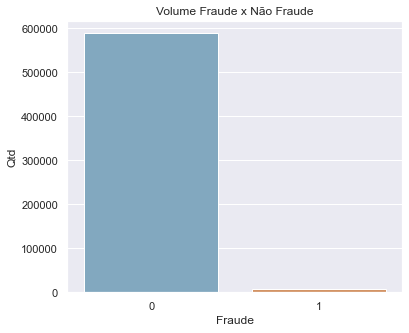
Os dados da ST passaram pela etapa de pré-processamento utilizando o ambiente de programação do Google Colab, o que constituiu em eliminar possíveis valores nulos e ajustar as configurações dos dados para atender os modelos propostos.

Os dados para o estudo após coletados foram processados para obtenção dos resultados utilizando a ferramenta Jupyter Notebook e a linguagem Python.  
Após esta etapa, foi verificado Informações como o tamanho do conjunto de dados e detalhes sobre as colunas, como o tipo de dados, e se possui valores nulos, neste caso não possuia. E já inicialmente verifiquei informações descritivas e contagens para algumas variáveis, como idade, categoria e fraude, para tratamento dos dados. **Foram** removidas colunas com um único valor (ZipCodeOri e ZipCodeMerchant), remoção de aspas em valores categóricos, substituição de valores "U" na coluna "idade", e formatação da coluna categoria.

Nessa etapa você começará a explorar seus dados de uma forma mais analítica, tentando elaborar ideias, levantar hipóteses e começando a identificar padrões em seus dados. Talvez você sinta a necessidade de voltar em passos anteriores, obter mais dados e tratá-los para conseguir responder ao problema proposto. Use e abuse de ferramentas estatísticas consistentes como testes de hipóteses, intervalos de confiança. Plote gráficos que te ajudem a obter insights interessantes: desde os mais simples até gráficos mais sofisticados como boxplots, mapas de calor, etc. Aqui o uso do Python e/ou R e suas poderosas bibliotecas gráficas (Matplotlib, Seaborn, ggPlot2, etc). Apresente trechos de código com as devidas justificativas.

Prosseguindo com o conjunto de dados, foi gerado um gráfico para melhor entendimento da variável alvo, onde fica claro que os dados possuem um desbalanceamento, onde os dados de fraude representam 1,21% dos dados totais, isso pode ser um precedente para a performance do modelo, portanto estes dados desbalanceados, serão trabalhados na etapa de feature enginerring, mais a frente.

Continuando com o conjunto de dados, um gráfico foi gerado para uma compreensão mais aprofundada da variável alvo. Tornou-se evidente que os dados apresentam um desbalanceamento, com as ocorrências de fraude representando apenas 1,21% do total, . Esse desbalanceamento pode resultar na baixa performance do modelo. Portanto, esses dados desbalanceados serão abordados na fase de feature engineering mais adiante.

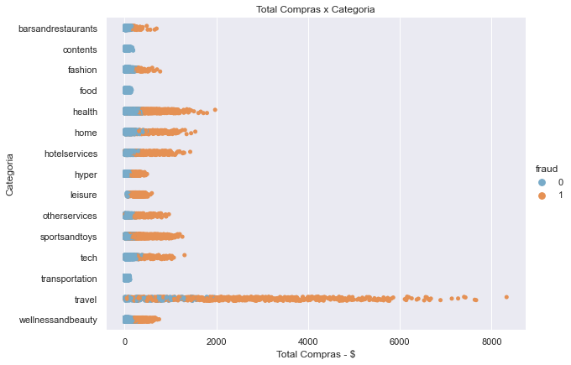
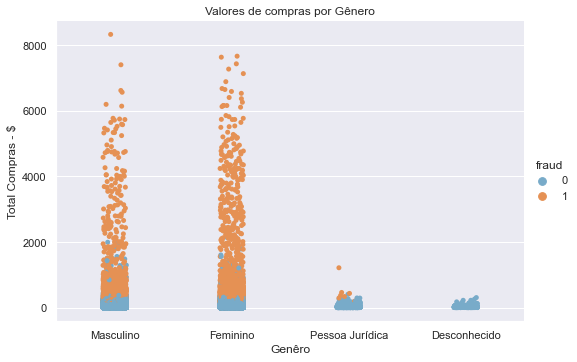


Fonte: Elaborado pelo autor (2023)

A partir disso foi separado o dataset de acordo com a coluna de fraude para facilitar na visualização e análise dos dados.   
Em seguida, foi feito a contagem da quantidade de transações pela coluna de categoria separados por fraude e nao fraude, adicionando uma coluna de total, e uma outra calculando o percentual de fraude em relação ao total de transações. E as percepções são de que apesar da classe de fraude terem um numero menor de transações, em algumas das categorias elas representam um alta quantidade. E em três categorias como transporte, comida e conteúdo em geral, não possui nenhuma transação de fraude.

A partir desse ponto, o conjunto de dados foi dividido com base na coluna de fraude para facilitar a visualização e análise dos dados.

Em seguida, foi a realizado a contagem da quantidade de transações para cada categoria, separadas entre fraude e não fraude e combinadas em um dataframe. Foram adicionadas duas colunas: uma para o total de transações e outra para calcular o percentual de fraudes em relação ao total de transações. As observações indicam que, embora a classe de fraudes tenha um número menor de transações, em algumas categorias elas representam uma quantidade significativa. Por outro lado, em três categorias - transporte, comida e conteúdo em geral - não há nenhuma transação de fraude.

Próximo passo foi iniciar uma verificação da dispersão de alguns dados categóricos, para entender a relação de algumas variáveis entre si, utilizando o enfoque da dispersão das transações de fraude e não fraude dentro das categorias.

Primeiramente avaliando o total de compras por categoria, nos possibilitando entender que os maiores gastos foram em compras fraudulentas, principalmente na categoria de viagem.  
Em seguida, verificando o total de compras por Gênero, mostrando que as maiores compras fraudulentas foi realizada pelo genero Feminino. O próximo gráfico sendo o total de compras por Idade, onde a classificação de 26-35 anos possui o maior gastos.

O próximo passo envolveu a análise da dispersão de alguns dados categóricos, visando compreender a relação de algumas variáveis entre si, com enfoque da dispersão entre transaçao de fraude e não fraude.

Primeiramente, foi avaliado o total de compras por categoria, o que nos permitiu perceber que os maiores gastos ocorreram em compras fraudulentas, especialmente na categoria de viagens. Em seguida, foi analisado o total de compras por gênero, evidenciando que as maiores compras fraudulentas foram realizadas pelo gênero feminino. O gráfico subsequente apresenta o total de compras por faixa etária, destacando que a faixa de idade de 26 a 35 anos registrou os maiores gastos.

Por ultimo, foi realizado uma visualização da dispersão considerando o dataset de dados de fraude, esse gráfico contém o total de compras por idade, e a dispersão por gênero. Podemos entender que a presença do genero feminino é mais forte nas classificações de 19-25, 26-35 e 56-65.

Por último, foi realizado uma visualização da dispersão utilizando o conjunto de dados de fraudes. O gráfico apresenta o total de compras por faixa etária, com a dispersão por gênero. É possível observar que a presença do gênero feminino é mais forte nas faixas etárias de 19-25, 26-35 e 56-65 anos.

Afim de verificar a distribuição dos dados, foi criado um grafico de barras com a soma do total de compras por categoria, separado por fraude e não fraude. E diferente da quantidade de transações fraudulentas por categoria, aqui nota-se que em relação a valor a categoria viagem se destaca, seguida de Saúde e então Esportes e Brinquedos.

Após a o passo acima, foi criado o grafico de boxplot, divido em 4 visualizações. As duas primeira em fraude e não fraude, onde é visto que a maior parte dos valores por transação de não fraude, não passaram de $500 doláres, exceto pela categoria viagem. E olhando pelo lado das transações que foram fraude, o maior gasto por transação foi também pela categoria viagem, e as outras categorias não ultrapassaram o gasto de $2000 doláres. Nas outras 2 visualizações, foi retirado a categoria de viagem, para conseguir uma melhor visualização das outras categorias. Tantos para transações normais quanto para de fraude, nota-se que existem bastante valores extremos, que fogem da sua distribuição normal.

Os proximos passos da análise, foi verificar a distribuição das frequências de cada variável categórica, com graficos de barras, separadas por transações de fraude e transações normais.

A fim de verificar a distribuição dos dados, foi criado um gráfico de barras apresentando a soma total de compras por categoria, separado entre fraude e não fraude. Diferentemente da quantidade de transações fraudulentas por categoria, observa-se que em relação ao valor, a categoria "viagem" se destaca, seguida por "Saúde" e, em seguida, "Esportes e Brinquedos".

Após essa etapa, foi criado um gráfico de boxplot dividido em quatro visualizações. As duas primeiras consideram transações de fraude e não fraude. Pode-se observar que a maioria dos valores por transação não fraudulenta não ultrapassou $500 dólares, exceto pela categoria "viagem". Ao analisar as transações fraudulentas, a categoria "viagem" também apresentou o maior gasto por transação, enquanto as outras categorias não ultrapassaram os $2000 dólares. Nas outras duas visualizações, a categoria "viagem" foi excluída para permitir uma melhor visualização das demais. Tanto para transações normais quanto para transações fraudulentas, é evidente a presença de muitos valores extremos que não se alinham com a distribuição normal.

Os próximos passos da análise envolveram a verificação da distribuição das frequências de cada variável categórica por meio de gráficos de barras, separados entre transações de fraude e transações normais.

Top of Form

Os próximos passos da análise envolveram a verificação adicional da distribuição das frequências de cada variável categórica por meio de gráficos de barras, separados entre transações de fraude e transações normais. Encontrando uma relação entre quantidade de transações x total de compras, onde por idade a faixa com a maior numero de transações continuou sendo na faixa etária de 26 a 35 anos, seguindo por gênero, onde o feminino também se destaca.   
E por ultimo foi analisado a frequencia de transações por comerciantes, onde o maior número de transações por comércio foi de 1634 transações, equivalente a 23% do número total de transações fraudulentas.

Os passos seguintes da análise focaram na exploração mais detalhada da distribuição das frequências de cada variável categórica, usando gráficos de barras separados entre transações fraudulentas e transações normais. Sendo entendido que existe uma relação entre a quantidade de transações e o total de compras. Observando-que, por faixa etária, a maior quantidade de transações continuou na faixa de 26 a 35 anos. Em relação ao gênero, o feminino também se destacou como o mais frequente.

Por último, realizou-se uma análise da frequência de transações por comerciantes. A maior quantidade de transações por comerciante foi de 1634 transações, representando aproximadamente 23% do total de transações fraudulentas.

~~foi realizado algumas plotagens para verificar a a distribuição dos valores gastos em diferentes categorias e em transações fraudulentas e não fraudulentas~~

criar gráficos de barras para verificar a distribuição de frequencias das variáveis categoricas.

1. **Visualização de Dados**: Várias visualizações gráficas estão sendo criadas para entender melhor os dados:
   * Um gráfico de contagem para fraude vs. não fraude.
   * Gráficos de caixa (boxplots) para visualizar a distribuição dos valores gastos em diferentes categorias e em transações fraudulentas e não fraudulentas.
   * Gráficos de barras para visualizar a quantidade de fraudes e não fraudes por categoria.
   * Gráficos de barras para quantidade de fraudes e não fraudes por categoria, separados.
   * Histogramas e gráficos de barras para idade e gênero em relação ao valor gasto.
   * Gráficos de dispersão para visualizar a relação entre o valor gasto e a categoria de transações, distinguindo entre fraudes e não fraudes.
2. **Análise Detalhada por Categoria**: Para cada categoria, estão sendo realizadas análises específicas incluindo a quantidade de fraude, a quantidade de não fraude, o total de transações e a porcentagem de fraudes em relação ao total.
3. **Análise por Idade e Gênero**: Diversas análises estão sendo realizadas para entender a relação entre idade, gênero e o valor gasto, tanto para fraudes quanto para não fraudes.
4. **Análise por Comerciante**: São criados gráficos de contagem para visualizar os comerciantes com mais ocorrências, tanto no geral quanto nas transações de fraude.
5. **Visualização de Dados com Filtros**: Algumas visualizações anteriores são repetidas com um filtro para mostrar apenas as transações relacionadas a comerciantes associados a fraudes.

Concluindo esta etapa de análise para identificar relação e padrões, será listado alguns insights obtidos:  
  
Desbalancemaneto de dataset - será testado na próxima etapa usando métodos de balanceamento.

Variavel Categoria - As categorias de transações parecem ter uma relação com as fraudes, e também que, está variavel indicar que determinadas categorias são mais propensas a fraudes do que outras.

Padroes entre transações - compras fraudulentas na categoria "Viagem" parecem ser especialmente elevadas. O gênero feminino parece estar associado a compras fraudulentas maiores e a faixa etária de 26 a 35 anos registra os maiores gastos.  
  
**Distribuição de Gastos -**  Muitas das transações de fraude têm valores extremamente elevados ao comparados com transações normais, que podem indicar atividades suspeitas

Essas questões serão abordadas nas próximas etapas de modelagem e análise preditiva.

Concluindo esta etapa de análise para identificar relações e padrões, foram obtidos os seguintes insights:

1. **Desbalanceamento do Dataset**: O desequilíbrio nos dados será abordado na próxima etapa usando métodos de balanceamento afim de melhorar o desempenho do modelo.
2. **Variável de Categoria**: As categorias de transações parecem estar correlacionadas com as fraudes. Algumas categorias podem ter maior propensão a fraudes do que outras.
3. **Padrões entre Transações**: As compras fraudulentas na categoria Viagem apresentam uma incidência particularmente alta. O gênero feminino parece estar associado a compras fraudulentas de maior valor, e a faixa etária de 26 a 35 anos tem os maiores gastos.
4. **Distribuição de Gastos**: Muitas transações fraudulentas têm valores excepcionalmente elevados em comparação com transações normais. Isso é um indicativo de atividades suspeitas e/ou anomalias.

Essas descobertas serão abordadas nas próximas etapas de modelagem e análise preditiva, onde métodos serão desenvolvidos para lidar com o desequilíbrio, e construir um modelo preditivo eficaz para detecção de fraudes.

Com base no texto fornecido, aqui estão algumas hipóteses e ideias que podem ser levantadas:

1. **Análise de Categorias**: As categorias de transações parecem ter uma relação com as fraudes. Algumas categorias, como "Esportes e Brinquedos" e "Saúde", têm uma quantidade significativa de transações fraudulentas, enquanto outras, como "Transporte", "Comida" e "Conteúdo em Geral", não têm transações de fraude. Isso pode indicar que determinadas categorias são mais propensas a fraudes do que outras.
2. **Compras por Categoria, Gênero e Faixa Etária**: O texto revela informações sobre os padrões de compras fraudulentas em relação à categoria, gênero e faixa etária. Por exemplo, compras fraudulentas na categoria "Viagem" parecem ser especialmente elevadas. O gênero feminino parece estar associado a compras fraudulentas maiores e a faixa etária de 26 a 35 anos registra os maiores gastos.
3. **Outliers e Distribuição de Gastos**: Há menção de outliers nos dados, especialmente em relação aos gastos por transação. Isso sugere que algumas transações, especialmente na categoria "Viagem", têm valores extremamente elevados que podem indicar atividades suspeitas.
4. **Relação entre Quantidade de Transações e Total de Compras**: Existe uma sugestão de que pode haver uma relação entre a quantidade de transações feitas por uma pessoa e o total de compras fraudulentas. A análise da frequência de transações por comerciantes também pode ser útil para entender os padrões de fraude.
5. **Modelagem e Análise Preditiva**: O texto não entra em detalhes sobre a modelagem ou análise preditiva realizada até agora. Pode-se supor que, após todas essas análises exploratórias, a próxima etapa seria a construção de modelos preditivos para detecção de fraudes, possivelmente utilizando técnicas como machine learning.
6. **Impacto nas Decisões de Negócios**: Compreender esses padrões de fraude pode ter implicações significativas para as decisões de negócios. A empresa pode usar essas informações para aprimorar seus sistemas de detecção de fraudes, ajustar estratégias de marketing ou tomar outras medidas para mitigar o risco de atividades fraudulentas.
7. **Necessidade de Mais Análises**: Embora o texto forneça uma visão geral das análises realizadas até o momento, ainda há várias questões que podem ser exploradas em maior profundidade, como a relação entre as variáveis categóricas e a variável alvo, a identificação de padrões temporais nas transações fraudulentas e a validação da eficácia de diferentes modelos de detecção de fraudes.

Em resumo, o texto fornece uma visão inicial das análises realizadas em um conjunto de dados de transações, focando na detecção de fraudes. As informações fornecidas indicam que foram feitas análises exploratórias para entender os padrões e relações nas transações fraudulentas e normais, além de levantar questões que podem ser abordadas nas próximas etapas de modelagem e análise preditiva.

Essas descobertas serão abordadas nas próximas etapas de modelagem e análise preditiva, onde métodos serão utilizados para lidar com o desequilíbrio entre classes, explorar ainda mais os padrões identificados, utilizando técnicas de codificação de variáveis para converter dados catégoricos em formatos númericos, permitindo a exploração da correlação entre os dados e por fim, desenvolver um modelo preditivo eficaz para detecção de fraudes.

Exatamente, essas descobertas serão tratadas nas próximas etapas de modelagem e análise preditiva. Para lidar com o desequilíbrio entre as classes, serão empregados métodos de balanceamento, como oversampling ou undersampling. Além disso, serão utilizadas técnicas de codificação de variáveis para converter os dados categóricos em formatos numéricos, permitindo a exploração da correlação entre os dados.

Com base nesses passos, o objetivo final será construir um modelo preditivo robusto e eficaz para a detecção de fraudes, que leve em consideração os padrões identificados, a relação entre as variáveis e os insights obtidos nesta fase de análise.

Para transformar dados categóricos em numéricos, você pode usar a técnica de codificação de variáveis categóricas. Uma maneira comum é usar a codificação one-hot (também conhecida como codificação de dummy), que cria colunas binárias para cada categoria. No entanto, se você tiver muitas categorias únicas, isso pode levar a um grande aumento no número de colunas. Outra abordagem é usar a codificação de rótulos (label encoding), que atribui um número inteiro para cada categoria.

Claro! O método **.corr()** é comumente usado em análise de dados e é usado para calcular a correlação entre variáveis em um conjunto de dados, geralmente representado em um DataFrame em bibliotecas como o pandas em Python. A correlação é uma medida estatística que indica a força e a direção da relação entre duas variáveis. Existem dois tipos comuns de correlação:

1. **Correlação de Pearson:** Isso mede a correlação linear entre duas variáveis contínuas. Varia de -1 (correlação negativa perfeita) a 1 (correlação positiva perfeita), com 0 indicando ausência de correlação. Um valor positivo indica que quando uma variável aumenta, a outra também tende a aumentar, e um valor negativo indica que quando uma variável aumenta, a outra tende a diminuir.
2. **Correlação de Spearman:** Isso mede a correlação monotônica entre duas variáveis, o que significa que pode identificar relações não-lineares. É útil quando os dados não seguem uma distribuição normal.

Exemplo de uso:

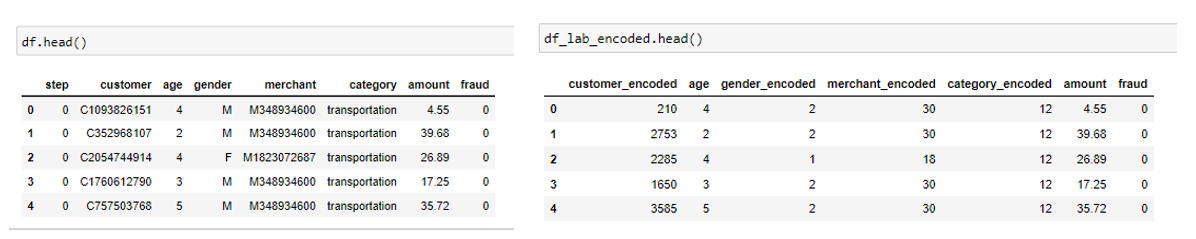
O próximo passo foi utilizar o método corr(), para calcular a correlação entre variáveis. A correlação é uma medida estatistica que indica a força e direção da relação entre duas variaveis. Onde um valor positivo indica que quando uma variavel aumenta a outra tambem tende aumentar, e um valor negativo indica que quando uma variavel aumenta a outra tende a diminuir. Observa-se que as variaveis que mais possuem correlação foi a de ‘Amount’ com ‘Fraude’ de 0,49%. As outras estatisticamente não apresentam relação significativa.

<figura?>

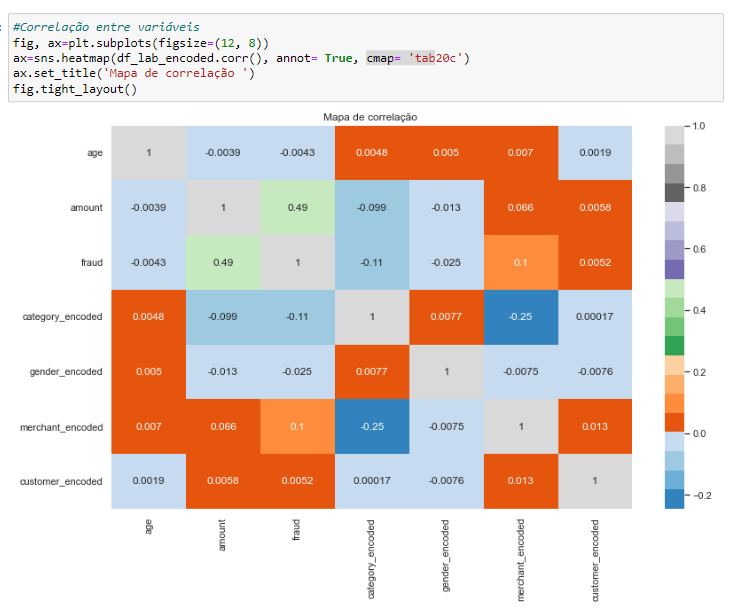
**Preparacao modelo**

Essa etapa se inicia com a preparação dos dados para serem utilizados no modelo de aprendizado de máquina. Codificar os dados é uma etapa importante, especialmente quando se trabalha com técnicas de desbalanceamento em problemas de classificação. A codificação é necessária para representar os dados categóricos em um formato numérico para que os algortimos possam entender. A técnica escolhida para este caso, foi a de ‘LabelEncoder’, ela transforma a variável categórica em numérica atribuindo um número inteiro a cada categoria. Cada categoria recebe um número único e sequencial.

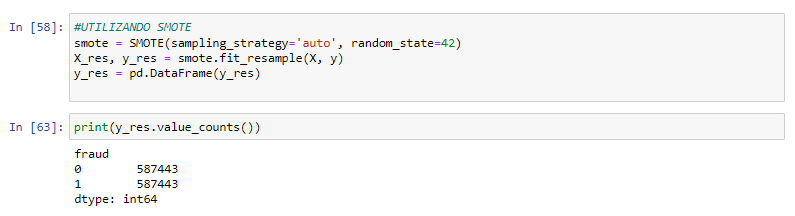
Abaixo um exemplo de como os dados estavam antes e como são apresentados depois da utilização da técnica de codificação.

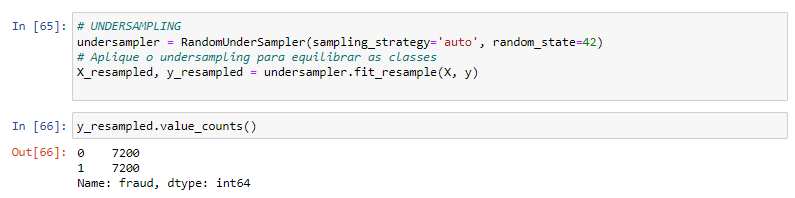


O próximo passo foi utilizar o método corr(), para calcular a correlação entre variáveis. A correlação é uma medida estatistica que indica a força e direção da relação entre duas variaveis. Onde um valor positivo indica que quando uma variavel aumenta a outra tambem tende aumentar, e um valor negativo indica que quando uma variavel aumenta a outra tende a diminuir. Observa-se que as variaveis que mais possuem correlação foi a de ‘Amount’ com ‘Fraude’ de 0,49%. As outras estatisticamente não apresentaram relação significativa.



Partindo para etapa de desbalanceamento, foi utilizado duas técnicas, é importante ressaltar que a aplicação dessa técnica é para garantir que o algoritmo não seja enviesado em direção à classe majoritária. A técnica de ‘SMOTE’, cria exemplos sintéticos da classe minoritária. Já a técnica ‘UNDERSAMPLING’, envolve a remoção de exemplos da maior classe para igualar o número de exemplos das duas classes. Isso reduz o tamanho do conjunto de dados, tornando-o mais equilibrado. No entanto, pode levar à perda de informações.





Independente dessas duas abordagens, a avaliação do desempenho do modelo é quem garantirá qual foi mais eficaz. O conjunto de dados foi dividido entre treino e teste, sendo 80% para treino e 20% para teste. Os primeiros modelos treinados estavam utilizando a técnica de Smote.

Para verificar os resultados obtidos através dos algoritmos e assim medir sua eficácia, é utilizado algumas métricas de desempenho, entre elas:

Acurácia (Accuracy): mede a proporção de previsões corretas em relação ao total de previsões. É a métrica mais simples e direta, mas pode ser enganosa em conjuntos de dados desbalanceados. Como foi tratado esse problema, ela pode ser informativa, mas ainda assim, não podemos enxerga-la sozinha.

Recall: Mede a capacidade do modelo de identificar corretamente todos os exemplos da classe positiva. É especialmente útil quando você deseja minimizar os falsos negativos.

Precisão (Precision): Mede a proporção de previsões corretas de verdadeiros positivos em relação ao total de previsões positivas. Útil para evitar falsos positivos.

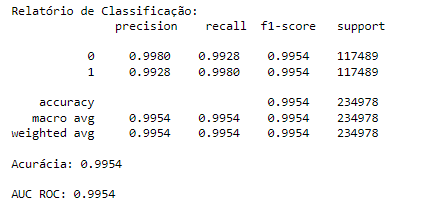
F1-Score: Combina a precisão e o recall em uma única pontuação, levando em consideração ambos os falsos positivos e falsos negativos. É útil para encontrar um equilíbrio entre precisão e recall.

AUC-ROC Score (Área sobre a curva): mede a capacidade do modelo de distinguir entre classes positivas e negativas. Quanto maior o AUC-ROC, melhor o modelo em discriminar entre as classes.

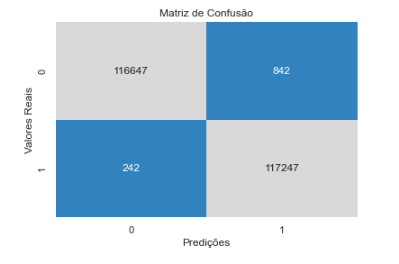
1° Modelo – RandomForestClassifier (Com SMOTE)

~~Foi criado primeiramente o modelo de aprendizado de máquina~~

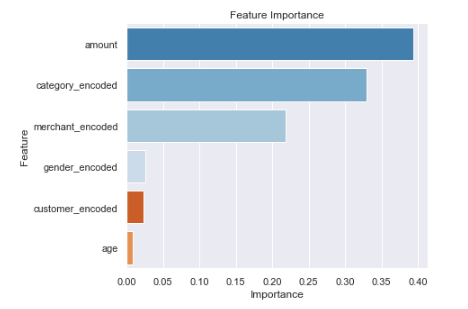
O modelo *‘RandomForestClassifier’* foi criado com parâmetro de n\_estimators= 100. Esse parametro define o numero de arvores de decisão na floresta. Quanto mais arvores, melhor o desempenho, porém maior o tempo para treinamento. E além desse, foi definido random\_state= 42 para garantir que toda vez que o código for executado o valor reproduzido será o mesmo. Os outros parametros foram mantidos o padrao do algoritmo.

~~A saída do modelo apresentou acuracia de 0.9954, precisao de 0.9928, recall de 0.9979, F1-score de 0.9953 e AUC-ROC de 0.9954.~~   
Abaixo o relatorio de classificação do primeiro modelo, onde fraude (1) e nao fraude (0). Este relatório traz a porcentagem de acerto de acordo com as classes, ele é interessante para ter uma visão mais completa e detalhada. E acompanhado do relatorio foi imprimido o valor da acuracia e da curva AUC ROC.

Os valores obtiveram um ótimo percentual de desempenho. E para visualizar as previsões feitas em números de transações, uma ferramenta interessante é a matriz de confusão, e olhando para essa matriz foram previstas 842 transações nao fraudulentas classificadas como fraude, e 242 transaçoes fraudulentas classificadas como não fraudulenta.



Foi utilizado uma outra técnica chamada ‘Feature Importance’, que auxilia no entendimento de quais variáveis tiveram o maior impacto na previsão do modelo, e é uma ferramenta útil que contribui na seleção das melhores features e também para otimização do modelo.



Os dados mostram que a variável amount obteve importancia de aproximadamente 40%, seguido de 32% de categoria e 22% de ‘merchant’.

2° modelo – XGBoost (Com Smote)

Este segundo modelo, utilizando ‘Extreme Gradient Boosting’ foi definido um parâmetro de random\_state= 42. E foi mantido o padrão que o algoritmo oferece.

Os resultados foram bem otimistas quantos ao primeiro modelo.

A matriz de confusão mostra que foram previstas 1.151 transações nao fraudulentas classificadas como fraude, e 537 transaçoes fraudulentas classificadas como não fraudulenta.

O grafico onde mostra a importancia das variaveis, mostra mais uma vez o valor obtendo o maximo de importancia, neste caso 65%, seguido de categoria com 18%.

3° modelo – RandomForest (Com Undersampling)

Os parametros foram mantidos o mesmo do primeiro modelo, porém utilizando a técnica de undersampling. Os dados mantiveram a mesma divisão que anteriormente de 80/20. A diferença maior é no total do conjunto de dados, que este possui um valor total de 7.200.

O desempenho obtido ainda assim foi bom, com uma acuracia de 97,3%, e um recall de 96% para classe de falsos negativos e 98% para classe de falsos positivos.

A matriz de confusão mostra 57 transaçoes como sendo falso negativo e 20 sendo falso positivo.

E a feature importance traz a categoria ‘amount’ com 65% de importancia.

4° modelo – XGboost (Com Undersampling)

O ultimo modelo, também utilizando a técnica de undersampling trouxe como resultado 97,5 de acuracia e uma precisao de 96,8% para a classe de falsos negativos. A matriz de confusão mostra 46 transaçoes não fraudulentas apontadas como fraude, e 26 transaçoes fraudulentas previstas como não fraude.

A variavel amount possui importancia de 49% seguida da de categoria de 33%.

O melhor conjunto de parâmetros é encontrado por meio de experimentação.